Biometrics 60, 427–435
June 2004

A Cautionary Note on the Robustness of Latent Class Models for Estimating Diagnostic Error without a Gold Standard

Paul S. Albert* **and Lori E. Dodd**

Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute,
Bethesda, Maryland 20892, U.S.A.
**email:* Albertp@ctep.nci.nih.gov

Summary. Modeling diagnostic error without a gold standard has been an active area of biostatistical research. In a majority of the approaches, model-based estimates of sensitivity, specificity, and prevalence are derived from a latent class model in which the latent variable represents an individual's true unobserved disease status. For simplicity, initial approaches assumed that the diagnostic test results on the same subject were independent given the true disease status (i.e., the conditional independence assumption). More recently, various authors have proposed approaches for modeling the dependence structure between test results given true disease status. This note discusses a potential problem with these approaches. Namely, we show that when the conditional dependence between tests is misspecified, estimators of sensitivity, specificity, and prevalence can be biased. Importantly, we demonstrate that with small numbers of tests, likelihood comparisons and other model diagnostics may not be able to distinguish between models with different dependence structures. We present asymptotic results that show the generality of the problem. Further, data analysis and simulations demonstrate the practical implications of model misspecification. Finally, we present some guidelines about the use of these models for practitioners.

Key words: Diagnostic accuracy; Latent class models; Misclassification; Prevalence; Sensitivity; Specificity.

1. Introduction

The lack of gold standard diagnostic truth often complicates evaluation of diagnostic errors of new medical tests. In some situations, gold standard evidence may be too costly to obtain, while in others, a method of diagnosing true status may not exist. A number of papers about estimating diagnostic error without a gold standard are in the literature. Hui and Walter (1980) proposed a methodology for estimating diagnostic error for binary tests under the assumption of conditional independence, i.e., when tests are independent conditional on disease status. Vacek (1985) examined the effect of conditional dependence on the estimation of diagnostic error when such conditional independence is assumed, and showed that parameter estimators are biased when independence is falsely assumed. Using an extensive set of simulations, Torrrance-Rynard and Walter (1997) also showed bias in this situation.

Several models that incorporate dependence across tests have been proposed. Espeland and Handelman (1989) developed a log-linear modeling approach for incorporating conditional dependence between tests in a latent class models framework. Qu, Tan, and Kutner (1996) proposed a random effects model that introduces conditional dependence between tests with a Gaussian random effect (GRE). Yang and Becker (1997) proposed a marginal approach that incorporates pairwise dependence across tests. Albert et al. (2001) proposed a finite mixture (FM) model for the dependence between tests and compared this model with others for estimating diagnostic error for a tumor marker. Hui and Zhou (1998) provided a review of much of this literature.

This note examines the robustness of inferences about diagnostic error to assumptions on the dependence structure between tests. We show that in many practical situations, estimators of sensitivity and specificity (as well as prevalence) are biased when the dependence structure is misspecified and that it is difficult to choose the correct dependence structure using likelihood comparisons and other model diagnostics. Our results support the conjecture of Begg and Metz (1990) that many models may fit the data equally well, but each provides different accuracy estimates. We focus on two particular situations throughout. The first considers multiple tests with the same diagnostic error when estimating a common sensitivity and specificity is of interest. This situation is common in diagnostic imaging studies to evaluate the diagnostic error of multiple raters scoring an image. In the second situation, tests have different diagnostic errors and interest focuses not only on sensitivity and specificity estimation, but also on ranking tests accordingly.

Section 2 reviews various latent class modeling approaches for estimating diagnostic error without a gold standard. In Section 3, we analyze a dentistry data set with these methods to illustrate the problem. We show the asymptotic properties of misspecification of the dependence structure in Section 4. Section 5 demonstrates the practical effects with a set of simulations. A discussion follows in Section 6 in which we provide guidelines for using these methods in light of our results.

2. Modeling Approaches

Let $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iJ})'$ be dichotomous test results for individual $i\,(i = 1, 2, \ldots, I)$, with Y_{ij} denoting the result from the *j*th of J tests. We denote d_i as the true unobserved disease status for patient i. All approaches assume a latent class model, where d_i is the latent class. The joint distribution of $\boldsymbol{Y}_i$ is

$$P(Y_{i1}, Y_{i2}, \ldots, Y_{iJ}) = \sum_{l=0}^{1} P(Y_{i1}, Y_{i2}, \ldots, Y_{iJ} \,|\, d_i = l)P(d_i = l), \qquad (1)$$

where $P(d_i = 1)$ is the disease prevalence which will be denoted by P_d. Under the assumption of conditional independence among tests (i.e., independence of tests given the true status), the probability of testing positive on test j given individual *i*'s true disease status is independent Bernoulli with probability $\rho_j(d_i)$. The sensitivity and specificity of test j are simply $\rho_j(1)$ and $1 - \rho_j(0)$, respectively. Hence, under conditional independence, $P(Y_{i1} = 1, Y_{i2} = 1, \ldots, Y_{iJ} = 1 \,|\, d_i) = \prod_{j=1}^{J} \rho_j(d_i)$. Various models allowing for conditional dependence have been developed which induce a positive dependence between tests (the conditional dependence structure is a complicated function of all the model parameters from each model except P_d). Qu et al.'s (1996) GRE model assumes that $(Y_{ij} \,|\, d_i, b_i)$ are independent Bernoulli with proportion given by $\Phi(\beta_{jd_i} + \sigma_{d_i} b_i)$, where the random variables b_i are standard normal and Φ is the cumulative distribution function (cdf) of a standard normal. Under this model $P(Y_{i1}, Y_{i2}, \ldots, Y_{iJ} \,|\, d_i) = \int \{\prod_{j=1}^{J} P(Y_{ij} \,|\, d_i, b)\}\phi(b)db$, where $\phi(b)$ is the standard normal density. In addition, under the GRE model, the sensitivity and specificity for the *j*th test are given by $\Phi(\beta_{j1}/(1 + \sigma_1^2)^{1/2})$ and $1 - \Phi(\beta_{j0}/(1 + \sigma_0^2)^{1/2})$, respectively. Similar models that describe dependence between tests through a continuous mixture model like the GRE model are possible. For example, when the diagnostic error is assumed to be the same across tests, we can assume $(Y_{ij} \,|\, d_i, p_{i,d_i})$ are independent Bernoulli with response proportion p_{i,d_i}, in which $p_{i,0}$ has a beta distribution with parameters α_0 and β_0 and in which $p_{i,1}$ has a beta distribution with parameters α_1 and β_1. Then, $(\sum_{j=1}^{J} Y_{ij}) \,|\, d_i = 0$ is beta binomial (BB) with parameters α_0 and β_0 (BB(α_0, β_0)) and $(\sum_{j=1}^{J} Y_{ij}) \,|\, d_i = 1$ is BB(α_1, β_1). Under this BB formulation, the sensitivity and specificity are simply $\alpha_1/(\alpha_1 + \beta_1)$ and $1/(\alpha_0 + \beta_0)$, respectively. In contrast with the GRE model, the BB model does not require numerical integration (such as Gaussian quadrature) for estimation or for evaluating the asymptotic bias.

Alternatively, Albert et al. (2001) proposed an FM model in which some individuals who are truly positive are always classified as positive by any test while others are subject to diagnostic error. Likewise, some truly negative subjects are always classified as negative by any test while others are subject to diagnostic error. In practice, this type of mixture model may occur in situations in which the most severely diseased patients and the healthiest patients are the easiest to correctly classify. Let l_{id_i} be an indicator of whether the *i*th subject, given disease status d_i, is always classified correctly, so that $l_{i1} = 1$ when a true positive subject is always positive and $l_{i0} = 1$ when a truly negative is always rated negative. Further, define $\eta_0 = P(l_{i0} = 1)$ and $\eta_1 = P(l_{i1} = 1)$. Test results Y_{ij} given d_i and l_{id_i} are independent Bernoulli with probability

$$P(Y_{ij} = 1 \,|\, d_i, l_{id_i}) = \begin{cases} 1 & \text{if } d_i = 1 \text{ and } l_{i1} = 1 \\ 0 & \text{if } d_i = 0 \text{ and } l_{i0} = 1 \\ \omega_j(1) & \text{if } d_i = 1 \text{ and } l_{i1} = 0 \\ 1 - \omega_j(0) & \text{if } d_i = 0 \text{ and } l_{i0} = 0, \end{cases} \qquad (2)$$

where $\omega_j(d_i)$ is the probability of the *j*th test making a correct diagnosis when the individual is subject to diagnostic error ($l_{i1} = 0$ or $l_{i0} = 0$). The FM model is closely related to the latent class model of Espeland and Handelman (1989) in which latent classes corresponding to unambiguously positive and negative cases are added. Under the FM model, the sensitivity and specificity of the *j*th test are $\eta_1 + (1 - \eta_1)\omega_j(1)$ and $\eta_0 + (1 - \eta_0)\omega_j(0)$, respectively.

Specific details about identifiability and estimation for the conditional independence, GRE model, and FM models are given within the reference corresponding to each of the approaches. To estimate test-specific diagnostic error, parameters of the conditional independence model are identifiable when $J \geq 3$. The parameters of the GRE and FM models are identifiable when $J \geq 4$. For estimating a common sensitivity and specificity across tests, $J \geq 5$ is required for the BB, GRE, and FM models. Estimation is based on maximizing $L = \prod_{i=1}^{I} \log P(Y_{i1}, Y_{i2}, \ldots, Y_{iJ})$ where $P(Y_{i1}, Y_{i2}, \ldots, Y_{iJ})$ is given by (1) with the joint distribution conditional on d_i depending on the model for the dependence structure. In results presented here, standard errors for model parameters were estimated using the bootstrap (Efron and Tibshirani, 1993).

3. Example

We analyzed Handelman's dentistry data (Handelman et al., 1986; Espeland and Handelman, 1989; Qu et al., 1996), in which five dentists (raters) evaluated 3869 dental X-rays for the presence of incipient caries (i.e., cavities that are beginning to appear), to illustrate the different methods for incorporating conditional dependence between tests (or, in this case, raters).

Initially, we fit models for estimating a common sensitivity and specificity across raters. Table 1 shows the frequency distribution of the sum of the number of positive responses for the five raters for this large group of patients. We present the expected frequency, log likelihood, chi-squared goodness-of-fit test, and estimates of the sensitivity, specificity, and prevalence for the conditional independence, FM, GRE, and the BB models. The expected frequencies for the conditional independence model differ substantially from the observed frequencies, where we observe many more all negative and all positive tests than predicted by the conditional independence model. The goodness of fit test for the conditional independence model is $\chi^2 = 18.56$, $df = 3$, which is highly significant and suggests the poor fit of the model. The FM, BB, and GRE models all have expected frequencies close to the observed frequencies. Log likelihoods and chi-squared values are

Table 1

Estimation of common sensitivity, specificity, and prevalence under the conditional independence (Indep), beta-binomial (BB), finite mixture (FM), and Gaussian random effects (GRE) models using Handelman's dentistry data

Positive		Expected frequency			
tests	Frequency	Indep	FM	BB	GRE
0	1880	1821.5	1879.5	1882.5	1880.4
1	1065	1132.9	1065.1	1058.8	1061.8
2	404	376.2	404.2	411.4	408.8
3	247	244.5	247.2	239.4	242.3
4	173	211.2	172.9	178.0	176.5
5	100	82.7	100.0	98.9	99.2
Total	3869				
$\widehat{\text{SENS}}$		0.658	0.645	0.518	0.457
		(0.017)[a]	(0.026)	(0.076)	(0.088)
$\widehat{\text{SPEC}}$		0.894	0.895	0.904	0.912
		(0.004)	(0.006)	(0.006)	(0.010)
$\widehat{P_d}$		0.166	0.169	0.240	0.294
		(0.010)	(0.017)	(0.063)	(0.073)
log L		−8726.5	−8717.7	−8718.0	−8717.8
χ^2		18.56	0.01	0.24	0.23
df		3	1	1	1

[a] Standard errors estimated using a bootstrap with 1000 bootstrap samples.

almost identical across the three models. In addition, none of the chi-squared goodness-of-fit tests are significant, suggesting the adequacy of the fit for all three models. Other model diagnostics are considered, including a comparison between model-based and observed marginal moments of test results for determining model adequacy proposed by Qu et al. (1996). We compared these model diagnostics for the FM and GRE models. The estimated marginal mean of a positive test and pairwise covariance between tests for both models are 0.197 and 0.044, respectively. These diagnostics are indistinguishable from each other and are essentially the same as the empirical marginal mean of 0.197 and pairwise covariance of 0.047. Interestingly, although all the measures of model adequacy are similar between models that allowed for conditional dependence, inferences on sensitivity and specificity as well as prevalence are different among the three models. For example, estimates of sensitivity are substantially smaller under the GRE model as compared with the FM model, while the standard errors are substantially larger under the GRE model.

Table 2 shows the estimates of rater-specific sensitivity and specificity for each of the five dentists for the conditional independence, FM, and GRE models. Although the rankings of sensitivity and specificity are nearly identical across each of the models, the estimated rater-specific sensitivity and specificity and their standard errors can be substantially different. The log likelihood for the conditional independence, FM, and GRE models are −7465.4, −7427.0, and −7421.8, respectively. The likelihoods for the FM and GRE models are relatively close and are substantially larger than the log likelihood for the conditional independence model.

To examine whether the difficulty of correct model identification is a problem asymptotically, in the next section we consider the large-sample behavior of these estimators.

4. Asymptotic Properties of Maximum Likelihood Estimators of Diagnostic Error

We investigate the asymptotic properties of estimators of diagnostic error when the dependence structure between tests is misspecified. The approach taken is similar to Heagerty and Kurland (2001) for examining bias in generalized linear mixed models. The misspecified maximum likelihood estimator for the model parameters, denoted by $\widehat{\boldsymbol{\theta}}^*$, converges to the value $\boldsymbol{\theta}^*$, where

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \mathrm{E}_T[\log L(\boldsymbol{Y}_i, \boldsymbol{\theta})], \tag{3}$$

and log$L(\boldsymbol{Y}_i, \boldsymbol{\theta})$ is the individual contribution to the log likelihood under the assumed model and the expectation is taken under the true model T. The notation

$$\mathrm{E}_T(\log L_M) = \mathrm{E}_T[\log L(\boldsymbol{Y}_i, \boldsymbol{\theta})]|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \tag{4}$$

denotes the expectation (taken under the true model T) of an individual's contribution to the log likelihood under the assumed model M when evaluated at $\boldsymbol{\theta}^*$. Sensitivity and specificity are model-dependent functional forms of the model parameters, SENS$^* = g_1(\boldsymbol{\theta}^*)$ and SPEC$^* = g_2(\boldsymbol{\theta}^*)$, where g_1 and

Table 2

Estimation of rater-specific sensitivity and specificity under the conditional independence (Indep), finite mixture (FM), and Gaussian random effects model (GRE) using Handelman's dentistry data

		Est. (SE[a])			Estimated ranks		
Rater		Indep	FM	GRE	Indep	FM	GRE
1	SENS	0.40 (0.026)	0.45 (0.038)	0.54 (0.120)	5	5	4
	SPEC	0.99 (0.002)	0.99 (0.003)	0.97 (0.013)	1	1	1
2	SENS	0.71 (0.025)	0.74 (0.034)	0.77 (0.100)	2	2	3
	SPEC	0.89 (0.007)	0.88 (0.008)	0.85 (0.026)	4	4	4
3	SENS	0.60 (0.028)	0.66 (0.040)	0.81 (0.190)	3	3	2
	SPEC	0.99 (0.003)	0.98 (0.005)	0.96 (0.021)	2	2	2
4	SENS	0.49 (0.022)	0.51 (0.026)	0.50 (0.060)	4	4	5
	SPEC	0.97 (0.005)	0.96 (0.007)	0.93 (0.022)	3	3	3
5	SENS	0.92 (0.014)	0.92 (0.018)	0.93 (0.070)	1	1	1
	SPEC	0.69 (0.011)	0.67 (0.012)	0.64 (0.032)	5	5	5

[a] Standard errors were estimated using a bootstrap with 1000 bootstrap samples.

Table 3

Large sample robustness of the assumed latent class beta-binomial (BB) model to the true dependence structure between tests. The true model is a finite mixture model (FM) with SENS = 0.75 and SPEC = 0.9 for differing P_d, η_0, and η_1.

Model parameters true model				Diagnostic error misspecified model		Expected log likelihood[a]	
η_0	η_1	P_d	J	SENS*	SPEC*	$E_{FM}[\log L_{FM}]$	$E_{FM}[\log L_{BB}]$
0.2	0.2	0.05	5	0.78	0.90	−1.82684	−1.82684
			6	0.64	0.90	−2.17092	−2.17127
			10	0.68	0.90	−3.52125	−3.52758
0.2	0.2	0.1	5	0.55	0.90	−1.98481	−1.98481
			6	0.53	0.90	−2.34536	−2.34586
			10	0.66	0.90	−3.74875	−3.75775
0.2	0.2	0.2	5	0.49	0.92	−2.24106	−2.24108
			6	0.52	0.92	−2.63077	−2.63158
			10	0.67	0.91	−4.13145	−4.14615
0.2	0.5	0.2	5	0.86	0.87	−2.10467	−2.10493
			6	0.88	0.86	−2.45631	−2.45745
			10	0.97	0.86	−3.82236	−3.83219

[a] Expected individual contribution to the log likelihood.

g_2 relate the model parameters to sensitivity and specificity. Estimators of sensitivity and specificity converge to SENS* and SPEC* under misspecified models.

We examine the asymptotic properties of estimators under departures from the FM model, although other approaches are possible. Specifically, we assume that the FM model is the correct formulation and calculate the expected log likelihood and asymptotic bias under various misspecified models.

Initially, we examine asymptotic bias with J tests assumed to have the same sensitivity and specificity. We evaluate the asymptotic bias of sensitivity and specificity estimators for different mixture model parameters as well as for different number of tests. The expected individual contribution to the log likelihood for the conditional independence model, BB model, and the FM model under the correctly specified FM model are given in the Appendix. In order to evaluate $\boldsymbol{\theta}^*$, the expected log likelihoods, defined in (3) and shown for specific models in the Appendix, are maximized using GAUSS using the Broyden, Fletcher, Goldfarb, and Shannon optimization method (Aptech, 1992).

Table 3 shows the asymptotic bias and expected individual contribution to the log likelihoods for the BB model when the true model is the FM model. The results show that estimators of sensitivity and specificity under a BB model can be severely biased when the true dependence structure is given by an FM model. Furthermore, our results suggest that it may be difficult to distinguish between the BB and the FM models with likelihood comparisons when we have small numbers of tests. For example, under an FM model with $\eta_0 = \eta_1 = 0.2$, $P_d = 0.1$, SENS =0.75, and SPEC = 0.9, the BB model estimates are SENS* = 0.55 and SPEC* = 0.90 for $J = 5$. The asymptotic bias for sensitivity is considerable, SENS−SENS* = 0.2. However, with only five tests, the expected individual contribution to the log likelihood for the correctly specified FM model is nearly identical (to five decimal places) to the expected individual contribution to the log likelihood for the BB model (evaluated at $\boldsymbol{\theta}^*$). Since the expected log likelihood is equal to the product of the number of individuals and the expected individual contribution to the log likelihood, a large number of individuals will be required to observe much of a difference when the expected individual log likelihoods are identical to the fifth decimal place. Thus, it may be impossible to distinguish between these two classes of models without extremely large samples. For a larger J, the two expected individual contributions to the log likelihoods are somewhat different, suggesting that, in this case, it may be possible to distinguish between the two classes of models with a larger numbers of tests in large samples. However, in most cases, designs with 10 or more tests are unrealistic.

Although the focus of this article is on estimating diagnostic error, we also found sizable asymptotic bias for prevalence estimation under a misspecified conditional dependence structure. For example, under an FM model with $\eta_0 = \eta_1 = 0.2$, $P_d = 0.1$, SENS = 0.75, and SPEC = 0.90, the estimator of prevalence under a misspecified BB model converged to 0.16, 0.16, and 0.12 for $J = 5$, 6, and 10, respectively.

In addition to the BB model, we investigate the asymptotic bias for the GRE model under a true FM model (table not provided). When $\eta_0 = \eta_1 = 0.2$, $P_d = 0.2$, SENS = 0.75, SPEC = 0.90, and $J = 5$, the GRE model has larger asymptotic bias than the BB model. For the GRE model, SENS* = 0.45 and SPEC* = 0.92 and the expected individual contribution to the log likelihood is −2.24106 which is very similar to the expected log likelihood of the BB model and is nearly identical to this quantity for the FM model (results for the BB and FM models are given in Table 3).

Next, we investigate the robustness of latent class models when we have four tests with different diagnostic error. In this case, interest focuses on evaluating the bias for individual test estimators of diagnostic error and in determining whether the rankings of sensitivities and specificities are preserved asymptotically. Table 4 shows the influence of the specification of the dependence structure between tests on accuracy estimation. There may be substantial bias in estimating sensitivity and specificity, but similar to our previous results, the expected log likelihood of the FM model (true model) and the expected

Table 4

Large-sample robustness of the Gaussian random effects (GRE) model assumption for four tests with different diagnostic errors when the true model is a finite mixture model (FM)

Model parameters				Diagnostic error					
True model				True model		Misspecified model		Expected log likelihood[a]	
η_0	η_1	P_d	Test	SENS	SPEC	SENS*	SPEC*	$\mathrm{E}_{\mathrm{FM}}[\log L_{\mathrm{FM}}]$	$\mathrm{E}_{\mathrm{FM}}[\log L_{\mathrm{GRE}}]$
0.5	0.5	0.2	1	0.80	0.95	0.73	0.95	−1.35814	−1.35814
			2	0.85	0.95	0.78	0.95		
			3	0.90	0.95	0.83	0.95		
			4	0.95	0.95	0.89	0.96		
0.9	0.1	0.5	1	0.80	0.95	0.78	0.82	−1.59592	−1.59593
			2	0.85	0.95	0.84	0.82		
			3	0.90	0.95	0.90	0.82		
			4	0.95	0.95	0.97	0.82		
0.5	0.5	0.5	1	0.80	0.95	0.72	0.85	−1.93643	−1.93662
			2	0.85	0.90	0.75	0.78		
			3	0.90	0.85	0.80	0.73		
			4	0.95	0.80	0.87	0.70		
0.5	0.5	0.1	1	0.80	0.95	0.58	0.92	−1.63991	−1.64034
			2	0.85	0.90	0.59	0.87		
			3	0.90	0.85	0.76	0.84		
			4	0.95	0.80	0.64	0.77		

[a] Expected individual contribution to the log likelihood.

log likelihood for the GRE model are nearly identical. Thus, it is nearly impossible, in many situations, to distinguish models based on likelihood comparisons *and* models are not robust to misspecification. In most cases, the rankings of sensitivity and specificity across tests are preserved asymptotically. However, there are exceptions as indicated in the last set of true model parameters given in Table 4 ($\eta_0 = \eta_1 = 0.5$ and $P_d = 0.1$), where the true sensitivity for tests 1–4 are 0.80, 0.85, 0.90, and 0.95, while SENS* are 0.58, 0.59, 0.76, and 0.64.

The results for model comparison in this section are asymptotic and are based on large-sample parameter estimates under the true or misspecified model. Of interest is determining whether correct model identification is a problem when the parameters under the correct or misspecified model are estimated in finite samples. Let $\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_I)$ and p be the number of parameters in the assumed model M. Let $\log L(\boldsymbol{Y}, \boldsymbol{\theta}) = \sum_{i=1}^{I} \log L(\boldsymbol{Y}_i, \boldsymbol{\theta})$. A Taylor series approximation can be used to express the expected log likelihood when $\boldsymbol{\theta}$ is estimated under the correctly specified model,

$$\mathrm{E}_{\hat{\boldsymbol{\theta}}}\{\mathrm{E}_T[\log L_T(\boldsymbol{Y}; \hat{\boldsymbol{\theta}})]\} = \mathrm{E}_T[\log L_T(\boldsymbol{Y}; \boldsymbol{\theta})] + \frac{1}{2}p + O\left(\frac{1}{I}\right). \qquad (5)$$

Under a misspecified model M where $\hat{\boldsymbol{\theta}}$ is calculated under M, the expected log likelihood is

$$\begin{aligned} &\mathrm{E}_{\hat{\boldsymbol{\theta}}}\{\mathrm{E}_T[\log L_M(\boldsymbol{Y}; \hat{\boldsymbol{\theta}})]\} \\ &\quad = \mathrm{E}_T[\log L_M(\boldsymbol{Y}; \boldsymbol{\theta})] \\ &\qquad + \frac{1}{2}\mathrm{tr}\left[\mathrm{Var}_T(\hat{\boldsymbol{\theta}})\mathrm{E}_T\left[\left.\frac{\partial^2 \log L_M(\boldsymbol{Y}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right]\right] \\ &\qquad + O\left(\frac{1}{I}\right). \end{aligned} \qquad (6)$$

Unfortunately, the term $\mathrm{Var}_T(\hat{\boldsymbol{\theta}})$ cannot be evaluated analytically. The difference in the penalty terms in evaluating the expected log likelihoods (5 and 6) may affect model choice in finite samples. We investigate this further with simulations in the next section.

5. Finite Sample Properties

We conducted a set of simulations studies to examine the finite sample properties for estimating a common sensitivity and specificity with different number of repeat tests and for estimating rater-specific sensitivity and specificity with four tests. Table 5 shows the results of an set of simulations in which we simulate according to both an FM model and a GRE model while estimating under both of these models. As with the asymptotic bias calculations, our simulation results show bias when estimating under the wrong model. Also, unless we have a large number of tests ($J = 10$), it is difficult to choose the correct model with high probability. For example, with a large sample size of $I = 1000$ and $J = 5$, the log likelihood for the FM model is larger than the log likelihood for the GRE model 74.3% of the time. However, the FM model has a larger log likelihood 86.0% of the time even when the true model is a GRE model. In addition, the log likelihoods for the two models are close when there are a small number of tests. For $I = 1000$ and $J = 5$, only in 11.2% and 10.8% of the time do the two log likelihoods differ by 1 or more when the true model is the FM and the GRE models, respectively (data not shown). We also find that, in large samples and for a small number of tests, the chi-squared goodness-of-fit test has low power to detect lack of fit for the incorrectly specified model. For example, when $I = 1000$ and $J = 5$, the probability of rejecting the GRE model when the true model is the FM model is only 4.9% and the probability of rejecting the FM model when the true model is the GRE model is only 3.4%.

Table 5

Simulations with a common sensitivity and specificity. Data were simulated either under the finite mixture model (FM) or the Gaussian random effects model (GRE) with $P_d = 0.2$, SENS = 0.75, and SPEC = 0.90. Data were generated under the FM model with $\eta_0 = \eta_1 = 0.2$ and under the GRE model with $\sigma_0 = \sigma_1 = 1.5$. Results are based on simulations with 1000 *realizations.*

			Avg. est.				χ^2 goodness-of-fit[a]			
		True	FM		GRE		% Reject			%
I	*J*	model	SENS	SPEC	SENS	SPEC	Indep	FM	GRE	$\log L_{\text{FM}} > \log L_{\text{GRE}}$
250	5	FM	0.75 (0.07)[b]	0.90 (0.02)	0.62 (0.17)	0.89 (0.03)	16.0	0.8	3.2	82.4
		GRE	0.84 (0.07)	0.94 (0.02)	0.72 (0.18)	0.89 (0.06)	95.1	0.2	4.6	88.4
250	6	FM	0.75 (0.05)	0.90 (0.01)	0.59 (0.18)	0.90 (0.03)	23.1	0.6	2.6	75.3
		GRE	0.84 (0.06)	0.94 (0.02)	0.72 (0.18)	0.89 (0.06)	99.6	0.9	1.8	61.5
250	10	FM	0.75 (0.03)	0.90 (0.01)	0.56 (0.17)	0.90 (0.02)	79.4	2.3	35.4	92.2
		GRE	0.83 (0.04)	0.94 (0.01)	0.73 (0.15)	0.89 (0.05)	100	50.2	3.6	9.0
1000	5	FM	0.75 (0.07)	0.90 (0.02)	0.62 (0.17)	0.89 (0.03)	88.7	0.1	4.9	74.3
		GRE	0.84 (0.07)	0.94 (0.02)	0.72 (0.18)	0.89 (0.06)	100	0.0	3.4	86.0
1000	6	FM	0.75 (0.03)	0.90 (0.01)	0.55 (0.13)	0.90 (0.01)	96.5	0.6	7.9	78.5
		GRE	0.83 (0.04)	0.94 (0.01)	0.73 (0.17)	0.88 (0.05)	100	5.8	1.1	36.8
1000	10	FM	0.75 (0.01)	0.90 (0.004)	0.54 (0.16)	0.90 (0.01)	100	2.3	98.7	99.6
		GRE	0.83 (0.02)	0.94 (0.01)	0.75 (0.13)	0.89 (0.04)	100	99.7	3.5	0.0

[a] Chi-squared goodness-of-fit test at the 0.05 significance level.
[b] Standard deviation of parameter estimates.

For a large number of tests, the probability of rejecting the wrong model with a chi-squared goodness-of-fit test is high, suggesting that we can correctly identify the model with a large number of tests.

We conducted a simulation in which we generated data according to an FM model with a large sample size ($I = 1000$) with differing sensitivity and specificity for four tests, and fit both the correctly specified FM model as well as the GRE model (Table 6). Estimation under the correctly specified FM model leads to estimators of diagnostic error that are unbiased and have small variability. Estimators under the misspecified model are highly biased and have large variability. As with our asymptotic calculations, it is difficult to identify the correct model. We identify the correct model in only 63% of the simulated realizations. The log likelihoods differ by a magnitude greater than 1 in only 24.3% of the simulated realizations. In addition, a chi-squared goodness-of-fit test rejects the GRE model in only 15.5% of the simulated realizations. We also examine the frequency in which the order in sensitivity and specificity is preserved. Under the correct model, the order is preserved in 85% and 87% of the simulations for sensitivity and specificity, respectively. This percentage is substantially reduced under the misspecified model (74% and 65%, respectively).

6. Discussion

In this note, we examine the robustness of latent class models to distributional assumptions about the dependence between tests. Although many authors have recognized that estimators of sensitivity and specificity may be biased when conditional independence is falsely assumed, there has been little discussion of the potential for bias when the dependence structure is incorrectly specified. Albert et al. (2001) showed through simulations that, for their particular design and parameter configuration, estimators of sensitivity and specificity were biased under a misspecified model. This current note demonstrates that the problem is more general. Of particular importance is that, not only are these estimators biased, but it may be nearly impossible to distinguish between different classes of models for the dependence structure under a realistic numbers of tests. Thus, in the case where we have few tests (enough for the correctly specified model to be identifiable), we need to make unverifiable modeling assumptions to make inference.

Although the required number of tests to distinguish between models will depend on the true model (and its parameter values) we found that it was very difficult to distinguish between models with only five or six tests, but substantially less difficult with 10 tests. Since in practice very few data

Table 6

Simulation with four tests in which test-specific sensitivity and specificity were estimated. Data were simulated under the finite mixture model (FM) with $\eta_0 = \eta_1 = P_d = 0.5$, with the sensitivity for the four raters being 0.80, 0.85, 0.90, *and* 0.95, *and with the specificity for the raters being* 0.95, 0.90, 0.85, *and* 0.80, $I = 1000$. *Results are based on* 1000 *simulated realizations.*

Test		Truth	Avg. est. FM	Avg. est. GRE
1	SENS	0.80	0.80	0.64
			(0.08)	(0.23)
	SPEC	0.95	0.95	0.79
			(0.03)	(0.11)
2	SENS	0.85	0.85	0.72
			(0.06)	(0.20)
	SPEC	0.90	0.90	0.72
			(0.05)	(0.10)
3	SENS	0.90	0.90	0.77
			(0.05)	(0.19)
	SPEC	0.85	0.85	0.68
			(0.06)	(0.11)
4	SENS	0.95	0.95	0.79
			(0.03)	(0.21)
	SPEC	0.80	0.80	0.64
			(0.08)	(0.13)
% [logL_{FM} > log L_{GRE}]			63	37
% reject χ^2_5 at 0.05 level test			2	18
% order preserved SENS			85	74
% order preserved SPEC			87	65

sets have as many as 10 tests, we feel that caution should be exercised in using these methods.

We primarily focused on using the likelihood in comparing models in this article. However, other approaches were also considered. For example, in the analysis of Handelman's dentistry data set, the chi-squared goodness-of-fit test, expected frequencies, and first two marginal moments of $\boldsymbol{Y}_i$ were all nearly identical for the BB, GRE, and FM models. In addition, our simulations (Tables 5 and 6) show that we often fail to reject the wrong model with a small number of tests using the chi-squared goodness-of-fit test.

Although our primary focus was on diagnostic error, we also found that prevalence estimation was biased under model misspecification. Under each model, the marginal mean can be expressed as P_d SENS + $(1 - P_d)(1 -$ SPEC). For Handelman's dentistry data (Table 1), the nearly doubling of the prevalence estimate under the GRE model as compared with the other models is consistent with the fact that marginal means are nearly identical across models and that sensitivity is substantially lower for the GRE model. In general, bias for the prevalence estimator is in the same direction as the bias for the sensitivity estimator and in the opposite direction as the bias for the specificity estimator.

Continuous mixture models such as the GRE model have been advocated in many settings (e.g., Hadgu and Qu, 1998). Our focus has, primarily, been on examining the robustness of continuous mixtures to FM alternatives. However, the problem of lack of robustness is a general problem for all conditional dependence models. For example, there is a lack of robustness for the FM model if the true model is a GRE model (see Table 5). Also, estimates of diagnostic error under Yang and Becker's (1997) model in which higher than second-order associations are set to zero, may be sensitive to the presence of higher order associations.

There is a recent literature on Bayesian approaches for making inference about diagnostic tests with dependent tests (e.g., Dendukuri and Joseph, 2001). A major advantage of the Bayesian approaches is the avoidance of identifiability problems with less than four tests by relying heavily on prior distributions for diagnostic error and prevalence. In situations in which good prior information is available, a Bayesian approach may help in distinguishing between models with few numbers of tests. This has been recognized by Black and Craig (2002) who proposed a Bayesian modeling averaging approach to remove some of the bias in model misspecification.

There has been an active debate over the validity of latent class models for estimating diagnostic error. This debate focuses on (i) whether a model-based consensus makes biological sense and (ii) whether latent class model parameters are identifiable with small numbers of tests (Alonzo and Pepe, 1999). Our results illustrate a further concern, namely that, even in cases when models are identifiable, their estimators may not be robust to the assumed dependence structure, and it may be impossible to distinguish between competing conditional dependence models with observed data using any measure of model adequacy.

The major point of our article is to caution the practitioner not to blindly apply methods for estimating diagnostic error without a gold standard. These results lead to a few recommendations. First, we emphasize the importance of conducting a gold standard test whenever possible. There is no substitute for this ideal design. Even collecting gold standard information on a fraction of subjects may aid in choosing a model. We are currently investigating approaches for such a design and believe that this type of hybrid design may provide a substantial improvement in robustness as well as in the ability to correctly choose from different models. Second, when a gold standard does not exist, a sensitivity analysis using very different methods for accounting for dependence should be conducted. This principle provided our motivation for studying the FM and GRE models, since each models the heterogeneity in a considerably different way. Although biological plausibility may aid the practitioner in favoring one model over another, a range of estimates from various models of diagnostic error (as well as standard errors) should be reported. Third, whenever possible, a large number of tests (or repeated tests) should be taken. This will improve the ability to distinguish between models for accounting for the dependence.

Acknowledgements

We thank Dr Margaret Pepe for her comments on an earlier version of this article. We thank the Center for Information

Technology, NIH, for providing access to the high-performance computational capabilities of the Beowulf cluster computer system. We thank the two reviewers and an associate editor for their constructive comments.

Résumé

Modéliser l'erreur diagnostique sans un ≪gold standard≫ est un domaine actif de la recherche biostatistique. Dans la majorité de ces approches, les estimations de la sensibilité, de la spécificité et de la prévalence basées sur ces modèles, dérivent d'une classe de modèles latents dans lesquels une variable latente représente un statut individuel vrai, mais non observé, vis-à-vis d'une maladie. Par souci de simplification, les approches initiales ont fait l'hypothèse que les résultats d'un test diagnostique pour un même sujet étaient indépendants sachant son vrai statut vis à vis de la maladie (c'est-à-dire hypothèse d'indépendance conditionnelle). Plus récemment, divers auteurs ont proposés des approches pour modéliser la structure de dépendance entre les résultats du test diagnostique sachant le vrai statut du patient. Ce travail aborde un problème important de ces approches. En particulier, nous montrons que, quand la dépendance conditionnelle est mal spécifiée, les estimateurs de la sensibilité, de la spécificité et de la prévalence peuvent être biaisés. Surtout, nous démontrons qu'avec un petit nombre de tests, des comparaisons de vraisemblance et d'autres modèles diagnostiques peuvent ne pas être capables de faire la distinction entre des modèles ayant des structures de dépendance différentes. Nous présentons des résultats asymptotiques qui montrent la généralité du problème. Plus loin (ensuite), des analyses de données et des simulations démontrent l'implication pratique de la mauvaise spécification des modèles. Enfin, nous présentons quelques guides d'utilisation de ces modèles.

References

Albert, P. S., McShane, L. M., Shih, J. H., et al. (2001). Latent class modeling approaches for assessing diagnostic error without a gold standard: With applications to p53 immunohistochemical assays in bladder tumors. *Biometrics* **57,** 610–619.

Alonzo, A. and Pepe, M. (1999). Using a combination of reference tests to assess the accuracy of a diagnostic test. *Statistics in Medicine* **18,** 2987–3003.

Aptech Systems. (1992). *Gauss Systems*, Version 3.0. Kent, Washington, D.C.: Aptech Systems.

Begg, C. B. and Metz, C. E. (1990). Consensus diagnoses and "gold standards." *Medical Decision Making* **10,** 29–30.

Black, M. A. and Craig, B. A. (2002). Estimating disease prevalence in the absence of a gold standard. *Statistics in Medicine* **21,** 2653–2669.

Dendukuri, N. and Joseph, L. (2001). Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics* **57,** 158–167.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* New York: Chapman and Hall.

Espeland, M. A. and Handelman, S. L. (1989). Using latent class models to characterize and assess relative-error in discrete measurements. *Biometrics* **45,** 587–599.

Hadgu, A. and Qu, Y. (1998). A biomedical application of latent class models with random effects. *Applied Statistics* **47,** 606–616.

Handelman, S. L., Leverett, D. H., Espeland, M. A., and Curzon, J. A. (1986). Clinical radiographic evaluation of sealed carious and sound tooth surfaces. *Journal of the American Dental Association* **113,** 751–754.

Heagerty, P. J. and Kurland, B. F. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika* **88,** 973–985.

Hui, S. L. and Walter, S. D. (1980). Estimating the error rates of diagnostic tests. *Biometrics* **36,** 167–171.

Hui, S. L. and Zhou, X. H. (1998). Evaluation of diagnostic tests without a gold standard. *Statistical Methods in Medical Research* **7,** 354–370.

Qu, Y., Tan, M., and Kutner, M. H. (1996). Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics* **52,** 797–810.

Torrance-Rynard, V. L. and Walter, S. D. (1997). Effects of dependent errors in the assessment of diagnostic test performance. *Statistics in Medicine* **97,** 2157–2175.

Vacek, P. M. (1985). The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics* **41,** 959–968.

Yang, I. and Becker, M. P. (1997). Latent variable modeling of diagnostic accuracy. *Biometrics* **53,** 948–958.

Received January 2003. *Revised November* 2003.
Accepted December 2003.

Appendix

Expected Log Likelihood for Independence, Beta Binomial, Gaussian Random Effects, and Finite Mixture Models When the True Model Is the Finite Mixture Model

We present the expected log likelihood for the misspecified and the correctly specified latent models under the assumption of a common sensitivity and specificity among J tests. Under the common sensitivity and specificity assumption, denote $\omega_0 = \omega_j(0)$ and $\omega_1 = \omega_j(1)$ for all j tests. Thus, the five parameters of the FM model are ω_0, ω_1, P_d, η_0, η_1. Let $S = \sum_{j=1}^{J} Y_j$. The expected log likelihood for a misspecified model under a true FM model can be written as

$$\begin{aligned}
&\mathrm{E}_{\mathrm{FM}}[\log(\boldsymbol{Y}, \boldsymbol{\theta})] \\
&\quad = \log\left[\sum_{i=0}^{1} P(S = 0 \mid d = i) P(d = i)\right] \eta_0 (1 - P_d) \\
&\qquad + \log\left[\sum_{i=0}^{1} P(S = J \mid d = i) P(d = i)\right] \eta_1 P_d \\
&\qquad + \sum_{a=0}^{J} \log\left[\sum_{i=0}^{1} P(S = a \mid d = i) P(d = i) \Big/ \binom{J}{a}\right] \\
&\qquad \times \binom{J}{J-a} \big[\omega_0^{J-a} (1 - \omega_0)^a (1 - \eta_0)(1 - P_d) \\
&\qquad + \omega_1^a (1 - \omega_1)^{J-a} (1 - \eta_1) P_d\big].
\end{aligned}$$

The vector $\boldsymbol{\theta}$ denotes parameters of the assumed model, with subscripts denoting the position in the vector. For all models, $\theta_1 = P(d = 1)$. For the independence model, $P(S \mid d = 1)$ and $P(S \mid d = 0)$ are binomial $(J,\ S,\ \theta_2)$ and binomial $(J,\ S,\ 1 - \theta_3)$, respectively. For the BB model, $P(S \mid d = 1)$ and $P(S \mid d = 0)$ are BB $(J,\ S,\ \theta_2,\ \theta_3)$ and BB $(J,\ S,\ \theta_4,\ \theta_5)$, respectively. For the GRE model as described in Section 2,

$$P(S \mid d = 1) = \int \binom{J}{S} \Phi(\theta_2 + \theta_3 b)^S \times \{1 - \Phi(\theta_2 + \theta_3 b)\}^{J-S} \phi(b)\, db \qquad \text{(A.1)}$$

$$P(S \mid d = 0) = \int \binom{J}{S} \Phi(\theta_4 + \theta_5 b)^S \times \{1 - \Phi(\theta_4 + \theta_5 b)\}^{J-S} \phi(b)\, db, \qquad \text{(A.2)}$$

where ϕ is the standard normal density, Φ is the normal cdf, and expressions (A.1) and (A.2) are evaluated using Gaussian quadrature with 50 quadrature points.

The expected log likelihood for the correctly specified FM model is

$$\begin{aligned}
&\mathrm{E}_{\mathrm{FM}}[\log_{\mathrm{FM}}]\\
&= \log\left[\left\{\eta_0 + (1-\eta_0)\omega_0^J\right\}(1-P_d) + (1-\eta_1)(1-\omega_1)^J P_d\right]\\
&\quad \times \left[\left\{\eta_0 + (1-\eta_0)\omega_0^J\right\}(1-P_d) + (1-\eta_1)(1-\omega_1)^J P_d\right]\\
&\quad + \log\left[\left\{\eta_1 + (1-\eta_1)\omega_1^J\right\}P_d + (1-\eta_1)(1-\omega_0)^J(1-P_d)\right]\\
&\quad \times \left[\left\{\eta_1 + (1-\eta_1)\omega_1^J\right\}P_d + (1-\eta_0)(1-\omega_0)^J(1-P_d)\right]\\
&\quad + \sum_{a=1}^{J-1}\left\{\log\left[(1-\eta_0)(1-\omega_0)^a\omega_0^{J-a}(1-P_d)\right.\right.\\
&\qquad\qquad \left. + (1-\eta_1)\omega_1^a(1-\omega_1)^{J-a}P_d\right]\\
&\qquad \times \binom{J}{a}\left[(1-\eta_0)(1-\omega_0)^a\omega_0^{J-a}(1-P_d)\right.\\
&\qquad\qquad \left.\left. + (1-\eta_1)\omega_1^a(1-\omega_1)^{J-a}P_d\right]\right\}.
\end{aligned}$$